Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

YOLOv8s_BSE: A Lightweight Multi-Scale Fusion Detection Model for Infrared Images of Substation Equipment

Weidong Zhang
Jilin University of Chemical Technology
Chen Su
Jilin University of Chemical Technology
Yunchang Dong
State Grid Jilin Electric Power Company Ltd.
Xiaoyong Bo
boxiaoyong@jlnku.edu.cn
JiLin Agricultural Science and Technology College
Ruyi Dong
Jilin University of Chemical Technology

Article

Keywords: YOLOv8s, Infrared images, Object detection, Lightweight, Multi-scale feature fusion

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9767556/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

YOLOv8s_BSE: A Lightweight Multi-Scale Fusion Detection Model for Infrared Images of Substation Equipment

Weidong Zhang[1,2], **Chen Su**[1,2], **Yunchang Dong**[3], **Xiaoyong Bo**[1,2,4,*], **and Ruyi Dong**[1,*]

[1]College of Electrical and Control Engineering, Jilin University of Chemical Technology, Jilin 132022, China
[2]Electrical and Information Engineering College, JiLin Agricultural Science and Technology College, Jilin 132101, China
[3]Electric Power Research Institute, State Grid Jilin Electric Power Company Ltd., Changchun 130021, China
[4]School of Electrical Engineering, Northeast Electric Power University, Jilin 132012, China
[*]Corresponding author@email: boxiaoyong@jlnku.edu.cn; dongruyi@163.com

ABSTRACT

With the rapid advancement of smart grid technologies, infrared image detection for substation equipment has become indispensable for early defect identification. Conventional approaches struggle to balance detection accuracy and real-time performance, primarily due to low contrast, strong noise, and blurred small-target features in infrared images. To overcome these limitations, we present YOLOv8s_BSE, a lightweight multi-scale fusion detection model built upon an enhanced YOLOv8s framework. Three key innovations are introduced: first, the dynamic C2f_Star module integrated with BiFPN establishes a cross-layer multi-scale feature interaction mechanism to enhance shallow feature complementarity; second, the efficient EMA channel attention module is incorporated into the neck to dynamically emphasize multi-scale features while suppressing background noise; third, the WIoUv3 loss function with an outlier assessment mechanism optimizes anchor quality evaluation, significantly improving small-target localization accuracy. Experimental results on our custom substation infrared dataset demonstrate that YOLOv8s_BSE achieves a mAP@0.5 of 96.8%, a 0.8% improvement over the baseline YOLOv8s. Notably, it reduces parameters to 9.61 million, cuts computational complexity by 7.7%, and decreases localization error by 12%. Our model provides a lightweight, high-precision solution for small infrared target detection in complex environments, facilitating the advancement of intelligent substation operation and maintenance.

Keywords: YOLOv8s; Infrared images; Object detection; Lightweight; Multi-scale feature fusion

1. Introduction

As the global energy system accelerates its transition to renewable energy sources, the large-scale grid integration of intermittent power such as wind and solar has significantly increased the operational pressure on power systems. Ensuring the reliability of substation equipment and realizing intelligent operation and maintenance have thus become critical pillars for safeguarding national energy security [1]. As core nodes of power transmission and distribution networks, substations' equipment health status directly determines the overall stability of the power grid, and equipment failure is one of the main causes of grid outages [2, 6]. The 2022 IEA World Energy Outlook report quantifies that approximately 42% of global power grid outages are attributable to substation equipment aging and malfunction, among which overheating is the most common direct failure mechanism, accounting for more than 60% of equipment-related incidents [2].

China, as the world's largest power consumer and renewable energy installer, faces particularly pressing demands for power grid intelligent upgrading. Domestic industry statistics show that in 2023 alone, more than 200 local power outages of 10kV and above occurred nationwide due to insufficient substation equipment monitoring, affecting over 1 million users cumulatively [3]. Traditional manual inspection methods have inherent limitations: the average inspection duration for a single substation ranges from 4 to 6 hours, and heavy reliance on technicians' subjective judgment results in a missed detection and false detection rate of 15%–20%, rendering them incapable of meeting the real-time monitoring requirements of large-scale power grids [3, 4].

Infrared thermal imaging technology, with its advantages of non-contact detection, uninterrupted power supply, and real-time monitoring capability, has become the mainstream technical approach for substation equipment condition monitoring. It enables early fault warning by capturing abnormal thermal signatures such as localized overheating and insulation degradation [4, 5]. However, substation infrared images inherently suffer from three major challenges: low contrast (temperature difference <5°C between targets and background, leading to >35% false positive rate for traditional threshold segmentation methods),

severe environmental noise interference, and blurred small-target features (critical components occupying only 3–10 pixels in long-distance imaging) [4, 5]. These limitations severely restrict the performance of automatic detection systems and increase the risk of cascading failures due to missed detections, underscoring the urgent need for technical advancements.

Deep learning-driven object detection has made remarkable progress in the field of substation infrared image analysis. Early studies predominantly concentrated on two-stage detection models, among which Faster R-CNN has been widely applied to power equipment defect detection. Wu et al. [7] embedded the InResNet module into Faster R-CNN to enhance feature representation capability, raising the mAP by 12.7% in infrared detection tasks. Nevertheless, this improved structure suffers from bulky model size and high computational cost, restricting its onboard deployment. To balance accuracy and efficiency, Ou et al. [8] proposed a lightweight modified Faster R-CNN tailored for substation infrared images, which reaches 93.2% mAP with an inference speed of 45 FPS. Even so, the model still struggles with feature extraction for dim and tiny targets under low-contrast and noisy infrared backgrounds, making it difficult to adapt to actual complex substation operating conditions.

As the most representative single-stage detection algorithm, the YOLO series has become the mainstream framework for substation infrared defect detection due to its excellent balance between accuracy and speed. Its iterative evolution has always focused on three core directions: multi-scale feature fusion, lightweight design and small target detection, which are exactly the key demands of substation infrared detection.

Liu et al. [9] first applied YOLOv4 (core improvement: PANet multi-scale feature fusion) to substation infrared detection, achieving an AUC of 0.91. Nevertheless, YOLOv4 lacks dedicated optimization for low-contrast thermal small targets, and its performance degrades significantly when the temperature difference is less than 5 °C. Wang et al. [10] introduced ECA-Net (core improvement: lightweight channel attention) into Ghost-YOLOv5, reducing the false positive rate to 2.8%. However, ECA cannot effectively suppress the complex electromagnetic interference and environmental noise unique to substations. Zhao et al. [11] combined diffusion model data augmentation with YOLOv6 (core improvement: anchor-free detection), raising the F1 score to 89%. Yet synthetic thermal data cannot fully replicate the real temperature gradient characteristics of on-site equipment. Deng et al. [12] proposed GhostNetV2-YOLOv7-tiny based on YOLOv7's lightweight improvement, which improved small-target accuracy by 19.3% with parameters below 1.2 million, but its cross-layer fusion ability is insufficient for densely arranged equipment.

Recent studies have focused on improving YOLOv8 for substation scenarios. Lu et al. [13] integrated coordinate attention into YOLOv8, achieving a 23.6% small-target accuracy gain, but did not optimize for weak thermal signals below 10 °C. Yang et al. [14] developed YOLOv8-MCDE with MobileNetV3 backbone, reducing parameters while maintaining small-target feature extraction ability, but neglected the shallow-deep feature complementarity required by infrared images. Sun et al. [15] replaced YOLOv8's backbone with EfficientViT, obtaining a 1.8% mAP improvement, but the performance gain is marginal. In addition, Shan et al. [16] proposed a lightweight SSD-based infrared detection network, but its inference speed is only 28 FPS, failing to meet real-time requirements. Zhai et al. [17] and Huangfu et al. [18] optimized YOLOv8 for UAV small-target detection, but these visible light-oriented methods cannot be directly migrated to infrared thermal imaging tasks.

Despite these advances, vanilla YOLOv8 still exhibits significant limitations in substation infrared detection scenarios. First, while the CSP-Darknet53 backbone improves feature reuse via cross-stage partial connections, it lacks sufficient sensitivity to subtle thermal signals with temperature differences below 10 °C. For targets smaller than 32 pixels, the limited convolutional receptive field impairs feature extraction, resulting in small-target recall rates generally below 70% [13, 14]. Second, the original PANet multi-scale fusion strategy fails to preserve spatial positional information during cross-scale feature propagation, weakening semantic correlations in complex environments (e.g., cloudy conditions, densely packed equipment) and increasing false positive rates by 15%–20% [15]. Third, conventional attention mechanisms such as SE [22] exhibit poor infrared noise suppression performance, with model performance degrading drastically when the dynamic range is compressed to 0.1–0.3. Furthermore, the baseline model's 11+ million parameters impose high computational overhead, limiting deployment on resource-constrained embedded devices [15, 17]. Consequently, while YOLOv8 performs well on general object detection tasks, it cannot simultaneously satisfy the stringent substation engineering requirements of mAP@0.5 > 95%, inference speed > 50 FPS, and parameter count < 10 million, highlighting a critical research gap.

To address the inherent defects of existing detection algorithms, this paper proposes an improved YOLOv8s_BSE model to meet the performance requirements of substation equipment infrared image detection. The main network improvement strategies are summarized as follows:

1) The backbone network integrates the cross-layer feature fusion module C2f_Star and reconfigurable convolution units. C2f_Star realizes high-dimensional feature mapping through star operation. Combined with the topological adjustment advantage of reconfigurable convolution, it strengthens the feature complementarity between shallow detail information and deep semantic information. Meanwhile, standard convolution is replaced by depthwise separable convolution, which greatly reduces computational complexity while maintaining feature extraction performance.

2) The neck network combines Bidirectional Feature Pyramid Network (BiFPN) with Efficient Channel Attention Module (EMA). BiFPN optimizes multi-scale semantic association through bidirectional feature propagation, while EMA realizes

adaptive channel weighting via a dual-branch structure, which effectively suppresses background interference and improves the signal-to-noise ratio of extracted features.

3) The detection head adopts an improved spatial pyramid pooling (SPPF) structure to expand the receptive field and enhance the ability to capture small target context information. Furthermore, a dynamic non-monotonic focal loss function derived from Wise-IoU is introduced. It adaptively adjusts gradient weights through an outlier evaluation mechanism, thereby effectively improving bounding box regression accuracy.

The remainder of this paper is organized as follows: Section 2 introduces the basic principle of the original YOLOv8s model; Section 3 elaborates the structural design and improvement details of the proposed model; Section 4 presents experimental settings, result verification and in-depth comparative analysis; Section 5 concludes the full research and gives future prospects.

2. Materials and Methods

2.1 YOLOv8s Framework

YOLOv8s, released by Ultralytics in 2023, is a state-of-the-art lightweight object detection model and a key milestone in the evolution of the YOLO series, which achieves a superior trade-off between detection accuracy, inference speed, and computational complexity [28]. As shown in Figure 1, its standard architecture consists of three modular components: the backbone feature extractor, the neck feature fusion module, and the decoupled detection head. The backbone extracts hierarchical multi-scale features from input images through successive downsampling operations, capturing both low-level edge details and high-level semantic information. The neck aggregates multi-level features via bidirectional top-down and bottom-up propagation to strengthen cross-scale feature correlation. The decoupled head separately performs object classification and bounding box regression, effectively reducing mutual interference between tasks.

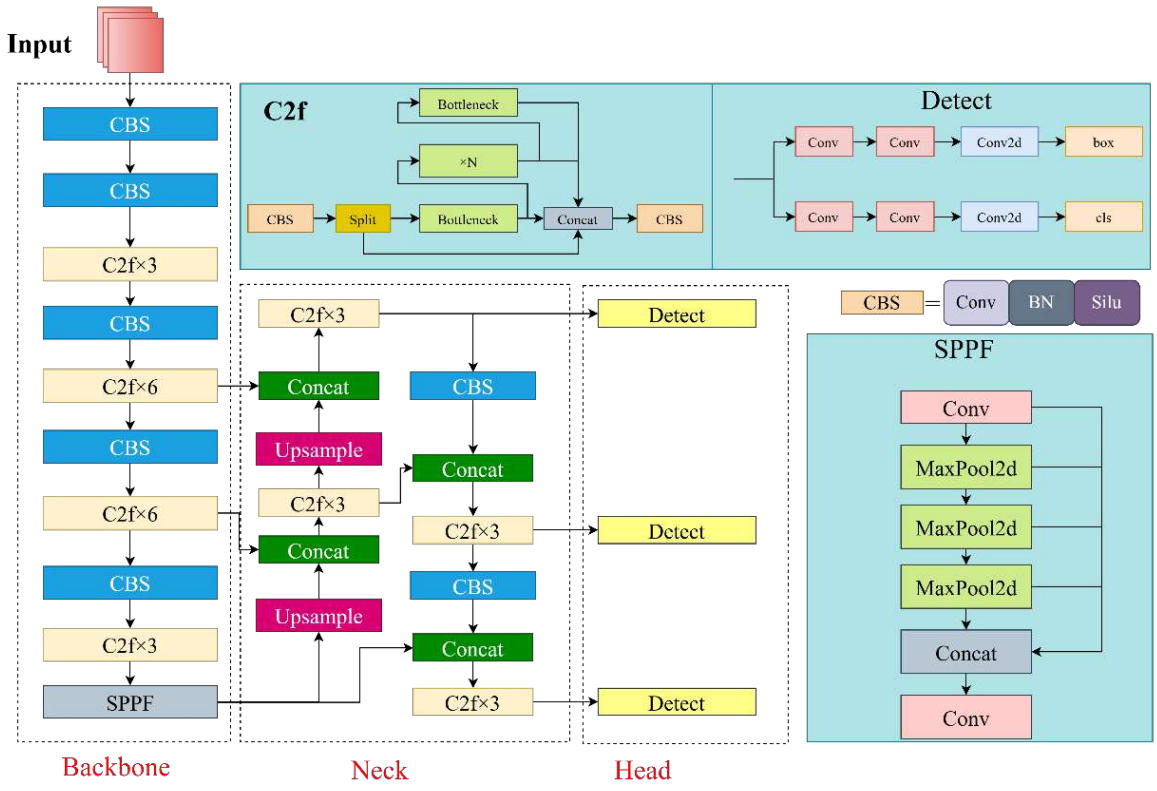


Figure 1. YOLOv8s network architecture diagram

The backbone network is built upon CSP-Darknet53, consisting of three core modules: the CBS module (convolution + batch normalization + SiLU activation) for basic feature transformation, the C2f module (inspired by YOLOv7's ELAN structure) for enhanced feature reuse, and the SPPF module for multi-scale contextual information fusion. The neck adopts the PANet structure, which combines top-down semantic propagation and bottom-up positional propagation to improve multi-scale detection performance. The detection head uses an anchor-free multi-branch design, with BCEWithLogitsLoss for classification and CIoU Loss for bounding box regression. These design choices enable YOLOv8s to maintain excellent performance on

resource-constrained edge devices [28], and it has become the most widely used baseline model for substation infrared detection in recent years [15].

To select the most suitable baseline model for substation infrared detection, we compared the performance of mainstream single-stage detection models. Among YOLO series variants, YOLOv5 exhibits insufficient small-target accuracy, while YOLOv7[19] and YOLOv9[20] have excessive computational overhead that limits embedded deployment. Within the YOLOv8 family, the n version has unsatisfactory detection accuracy for weak thermal targets, while the m/l/x versions have too many parameters for edge device deployment. Other state-of-the-art models such as RT-DETR and YOLOv10 have higher inference latency on low-power embedded platforms. In contrast, YOLOv8s achieves 56.8% mAP@0.5 and 160 FPS inference speed on the COCO dataset, achieving the best balance between accuracy and efficiency. Therefore, this study selects YOLOv8s as the baseline model for subsequent improvements. However, the original YOLOv8s still has three key limitations in substation infrared detection scenarios: (1) insufficient sensitivity to weak thermal signals with temperature differences below 10°C; (2) poor cross-scale feature fusion ability for densely arranged small targets; (3) weak suppression of complex electromagnetic and environmental noise. These limitations directly motivate the three targeted improvements proposed in this paper.

2.2 YOLOv8s_BSE model

YOLOv8s_BSE is an enhanced lightweight model tailored for substation equipment infrared image detection. Built upon the efficient architecture of YOLOv8s, it improves small-target detection performance through three targeted innovations. First, the C2f_Star cross-layer feature fusion module is embedded into the backbone network. This module incorporates the star operation to realize dynamic feature fusion, mapping input features to a high-dimensional space via element-wise multiplication. This significantly enhances the sensitivity to weak thermal signals while maintaining parameter efficiency. Second, the BiFPN-EMA hybrid mechanism is integrated into the neck network. BiFPN adaptively weights the contributions of features at different resolutions through bidirectional cross-scale connections and fast normalized fusion, while EMA attention recalibrates multi-scale channel features via a dual-branch structure, effectively suppressing complex background interference. Finally, the detection head adopts the WIoUv3 loss function, which refines bounding box regression through a dynamic non-monotonic focusing mechanism to improve small-target localization accuracy. The overall architecture is illustrated in Figure 2. With clearly defined hierarchical modules, it forms a progressive optimization pipeline of "feature enhancement–multi-scale fusion–precise localization", which collectively addresses the core challenges of low infrared image contrast, blurred small-target features, and complex background interference.

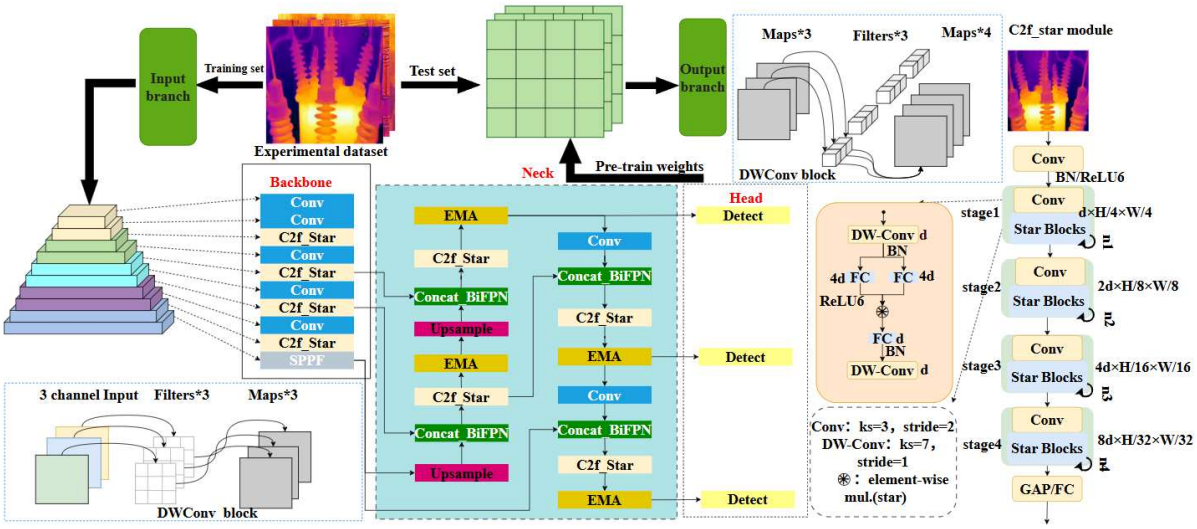


Figure 2. Structural framework of YOLOv8s_BSE

2.2.1 Lightweight Improved C2F Module Embedded with StarNet

Conventional CNNs, however, struggle to extract fine thermal features from substation infrared images—these images typically have low contrast and small, blurred targets, and standard convolutions lose detailed information due to their limited receptive fields. StarNet [21] addresses this limitation via its core Star Operation, which maps features into a high-dimensional implicit space. Its hierarchical architecture is structured as follows: Stages 1–2 adopt standard convolution and downsampling to extract low-level texture features, Stages 3–4 stack Star Blocks for feature enhancement and fusion, and cross-stage connections use depthwise separable convolution to expand channels progressively.

As an efficient feature mapping approach, the Star Operation projects input features into a high-dimensional representation space via element-wise operations, offering three notable advantages over conventional convolutions. First, its core mathematical definition is:

$$Star(X) = \text{act}(W_1^T X) \odot \text{act}(W_2^T X) \tag{1}$$

Here, X denotes the input feature map, $W_1, W_2 \in \mathbb{R}^{C \times C}$ denote two independent learnable weight matrices, $\text{act}(\cdot)$ denotes the ReLU6 activation function, and $\odot$ denotes element-wise multiplication (i.e., the star operation).

This dual-path design enables implicit high-dimensional feature mapping without increasing the network width. Expanding Equation (1) yields:

$$Star(X) = \left(\sum_{i=1}^{d} w_1^i x^i \right) \odot \left(\sum_{j=1}^{d} w_2^j x^j \right) \tag{2}$$

Here, d represents the channel dimension of the input features. From the above definition, we can derive the dimension of the linearly independent features generated by the star operation as follows:

$$D = \frac{(d+2)(d+1)}{2} \approx \left(\frac{d}{\sqrt{2}} \right)^2 \tag{3}$$

Theoretical analysis demonstrates that the star operation can generate quadratic-order high-dimensional features in a low-dimensional computational space, significantly improving the discriminative capability for low-contrast thermal features.

Furthermore, the star operation requires only one-quarter of the parameters of a standard convolution with the same input and output dimensions. The parameter quantity comparison formula is as follows:

$$Params_{Star} = 2 \times C^2 = \frac{1}{4} Params_{Conv}(3 \times 3) \tag{4}$$

In this context, $\text{Params}_{Conv}(3 \times 3) = 9C^2$ denotes the parameter quantity of a standard 3×3 convolution. This characteristic makes the star operation particularly suitable for embedded deployment scenarios in substations with limited computational resources. Third, the feature representations generated by the star operation exhibit stronger nonlinearity, enabling more effective differentiation between weak target features and complex background noise.

The star operation is embedded into the C2f module of YOLOv8s to develop the proposed C2f_Star module. This module employs a dual-branch parallel architecture: the main branch preserves the original feature path to retain fine-grained spatial details, while the auxiliary branch incorporates the star operation to perform implicit high-dimensional feature mapping. The outputs of the two branches are fused via element-wise addition through a residual connection, followed by batch normalization to ensure training numerical stability. The complete mathematical formulation of the C2f_Star module is presented below:

$$C2f_Star(X) = BN\Big(X \oplus \big(act(W_1^T \bullet Conv_{1\times1}(X)) \odot (W_2^T \bullet Conv_{1\times1}(X)) \big) \Big) \tag{5}$$

Here, $Conv_{1\times1}$ denotes a 1×1 convolutional layer for channel dimension adjustment; $\oplus$ denotes residual fusion; and BN denotes batch normalization, which ensures training numerical stability.The C2f_Star module consists of four core components: a 1×1 convolutional layer, a star operation layer for high-dimensional nonlinear mapping, a feature fusion layer for multi-scale feature integration via skip connections, and an output normalization layer.These components form an improved basic unit that achieves a balance between lightweight design and feature enhancement capability.

2.2.2 EMA Mechanism Based on Channel Weight Allocation Strategies

Attention mechanisms have become a core component for enhancing feature discriminability in object detection. The Squeeze-and-Excitation Network (SENet) improves network representation capability through automatic channel-wise recalibration [22]. However, SENet primarily focuses on channel relationships while neglecting spatial positional information. To address this limitation, the Coordinate Attention (CA) mechanism was proposed, which embeds positional information into channel attention to significantly boost detection performance [23]. Nevertheless, CA fails to capture long-range interactions between horizontal and vertical dimensions, and its 1×1 convolution has a limited receptive field, which hinders local cross-channel information communication.

The Efficient Multi-scale Attention (EMA) mechanism fully inherits the advantages of CA while effectively overcoming its limitations. It establishes cross-channel relationships without reducing channel dimensions, thereby avoiding the adverse effects that may arise during deep visual feature extraction [24].

Substation equipment infrared detection requires separating weak target signals from complex thermal background noise. The EMA mechanism [25] addresses this challenge with a lightweight dual-branch design: one branch performs channel-wise feature recalibration without changing dimensions, while the other captures spatial interaction features in parallel. One branch extracts global contextual information, while the other captures local detailed features. Cross-dimensional feature interaction then generates joint channel-spatial attention weights, outputting an enhanced feature map with the same spatial dimensions as the input, as shown in Figure 3.

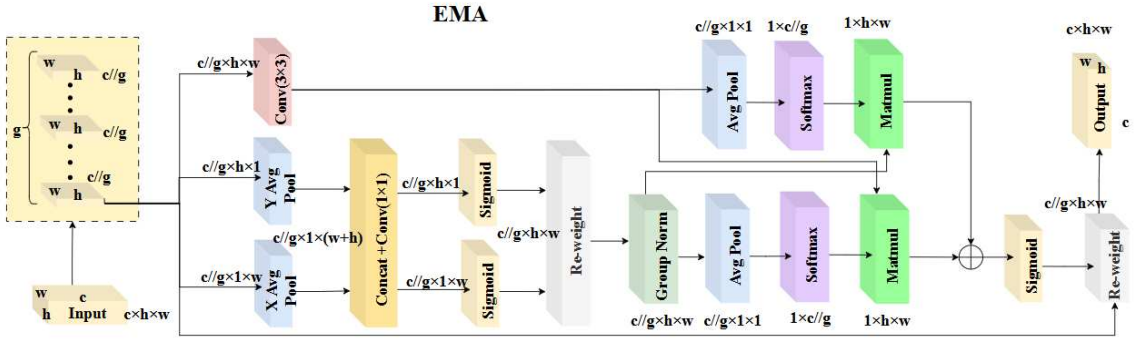


Figure 3. Schematic diagram of the Efficient Multi-scale Attention module

The EMA module adopts a feature grouping design for input processing. For any input feature map $X \in \mathbb{R}^{b\times C\times H\times W}$, where b is the batch size, C is the number of input channels, and H and W represent the height and width of the input feature map, respectively. The module first splits it into G sub-features along the channel dimension, namely $X = [X_0, X_1, ..., X_{G-1}]$, among which $X_i \in \mathbb{R}^{(C/G)\times H\times W}$ and $G \leq C$. Then reshape the grouping dimension into a batch dimension:

$$X_g = \text{reshape}(X, (bG, C/G, H, W)) \tag{6}$$

This operation reduces the input channel count of convolutional layers from C to C/G, significantly lowering computational overhead without incurring any information loss from channel dimension reduction.

The first subnetwork employs a 1×1 convolution design inspired by coordinate attention: it performs adaptive average pooling along both height and width dimensions to extract global spatial cues from grouped features. The two directional pooling outputsare concatenated, processed by a dimension-preserving 1×1 convolution and sigmoid activation to generate spatially-aware channel weights. We multiply these weights element-wise with the input features, followed by group normalization.

$$X_1 = GN\left(X_g \odot \sigma(X_h) \odot \sigma(X_w)^T\right) \tag{7}$$

Where $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$ denotes the sigmoid function, X_h and X_w represent the positional encoding features in the height and width directions, respectively. The second subnetwork consists of a 3×3 convolutional branch, which performs 3×3 convolutions on the grouped features to capture local multiscale spatial structural information:

$$X_2 = Conv_{3\times 3}(X_g) \tag{8}$$

The cross-spatial learning mechanism is the core innovation of EMA. The outputs of the two branches are each subjected to global 2D average pooling, and global feature vectors are generated via the Softmax function. These vectors are then multiplied by the spatial features of the other branch to produce two complementary spatial attention maps:

$$A = matmul\left(Z_1, reshape(X_2, (bG, C/G, HW))\right) + matmul\left(Z_2, reshape(X_1, (bG, C/G, HW))\right) \tag{9}$$

In particular, Z_1 and Z_2 denote the global feature vectors obtained by applying global average pooling and Softmax to X_1 and X_2, respectively. The attention weights are passed through a Sigmoid activation function and multiplied element-wise with the grouped input features, then reshape them back to the original dimension to obtain the final enhanced feature output.

$$Y = reshape\left(X_g \odot \sigma(A), (b, C, H, W)\right) \tag{10}$$

This output has the same dimensions as the input feature maps and can be directly stacked into the existing network architecture.

2.2.3 The BiFPN Network Based on Bidirectional Feature Propagation and Multiscale Fusion

Infrared images of substation equipment contain multi-scale objects, ranging from large cabinets to medium-sized components and tiny connectors. Effective cross-scale feature integration is the key to improving detection accuracy. The traditional Feature Pyramid Network (FPN) [24] only uses a top-down path for feature propagation, while the Path Aggregation Network (PANet) adds a bottom-up path but fails to distinguish the relative contributions of different input features; neither approach can achieve efficient multi-scale feature interaction. The Bidirectional Feature Pyramid Network (BiFPN) [26] is based on bidirectional cross-scale connections and fast normalized weighted fusion, enabling balanced transmission of deep semantic information and shallow detailed positional information, thus meeting the multi-scale detection requirements of substation equipment infrared images.

BiFPN introduces three key improvements to cross-scale connections. First, it removes nodes with only one input—no fusion occurs there, so computation is wasted. Second, it adds skip connections between same-level inputs and outputs, preserving original feature information at no extra cost. Third, it constructs iterative bidirectional (top-down and bottom-up) paths for stronger multi-scale fusion. Since features at different resolutions contribute unevenly, simple summation dilutes useful signals; BiFPN instead assigns learnable weights to each input and applies fast normalization to maintain training stability. Compared with Softmax-normalized fusion, this strategy accelerates GPU inference by 26%–31% with an accuracy loss of less than 0.1% [26].

Building upon the BiFPN fast normalized weighted fusion framework, we propose an improved Concat_BiFPN feature fusion module. Compared with the original BiFPN, this module innovatively adopts a weighted channel concatenation fusion strategy, fully preserving the independent information of each input feature; meanwhile, it implements runtime dynamic weight initialization, eliminating the need to predefine weight dimensions, thus significantly enhancing the module's portability and scalability. The overall structure of Concat_BiFPN is shown in Figure 4.

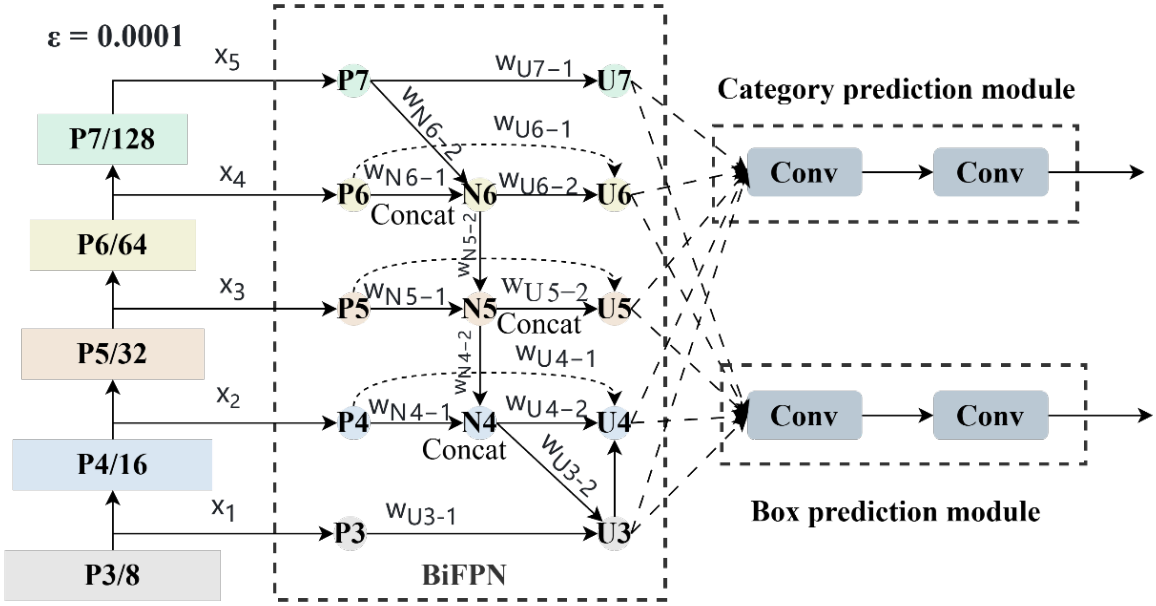


Figure 4. Schematic diagram of the Concat_BiFPN feature fusion module

The module first performs fast normalization on the trainable weights, as shown in Equation (11):

$$w_i' = \frac{w_i}{\sum_{j=1}^{N} w_j + \varepsilon} \tag{11}$$

where w_i is the initial weight of the i-th input feature; N is the total number of input features; $\varepsilon = 0.0001$ is a small constant used to prevent numerical anomalies caused by division by zero. The normalized weight w_i' ranges from 0 to 1 and can be automatically optimized via gradient descent during training, adaptively matching the contribution of features at different scales.

After weight normalization, the module applies weighted processing to each input feature and then concatenates them along the channel dimension for output. The fusion calculation is given by Equation (12):

$$O = \text{Concat}(\omega_1' \cdot x_1, \omega_2' \cdot x_2, ..., \omega_N' \cdot x_N) \tag{12}$$

where x_i denotes the input feature of the i-th layer, and $\text{Concat}(\cdot, d)$ represents the concatenation operation along the d-th dimension. In this work, $d = 1$ corresponds to the channel dimension. Compared with the element-wise addition fusion method of the original BiFPN, channel concatenation can preserve the independent information of each input feature, avoid information loss caused by feature superposition, and is more suitable for capturing weak thermal anomaly features in infrared images.

2.2.4 A Bounding Box Loss Optimization Mechanism Based on Wise-IoU Dynamic Non-Monotonic Focusing

Bounding box regression loss directly determines the localization accuracy of object detection models. In substation equipment infrared images, small targets such as tiny connectors exhibit weak thermal signatures, and training data inevitably contains low-quality samples, including those with annotation errors or missing labels. Traditional loss functions such as CIoU and SIoU fail to distinguish sample quality differences; blindly emphasizing the regression of low-quality samples introduces harmful gradients, leading to degradation of the model's generalization ability. Although Focal-EIoU introduces a non-monotonic focusing mechanism, it uses a static quality threshold and cannot dynamically adjust the gradient allocation strategy during training.

WIoUv3 (Wise-IoUv3) [27] is adopted as the bounding box regression loss function. This loss function is constructed on the basis of a dynamic non-monotonic focusing mechanism, which adaptively allocates gradient gains according to the outlier degree of anchor boxes while concurrently down-weighting both high-quality and low-quality samples. This mechanism enables the model to concentrate on learning from moderate-quality samples, thereby effectively isolating the training process from interference caused by annotation noise.The computational procedure of WIoUv3 comprises three core steps. First, the intersection-over-union (IoU) between the predicted bounding box and the ground-truth bounding box is computed to derive the baseline IoU loss:

$$L_{bU} = 1 - IoU \tag{13}$$

where IoU refers to the ratio of the intersection area to the union area between the predicted bounding box and the ground truth box.

An exponential moving average is adopted to update the running mean of IoU loss and serve as a dynamic quality benchmark:

$$\bar{L}_{IoU}^{(t)} = (1-m) \cdot \bar{L}_{IoU}^{(t-1)} + m \cdot E[L_{bU}^{(t)}] \tag{14}$$

where $\bar{L}_{IoU}^{(t)}$ denotes the running mean of IoU loss at the t-th batch, and m represents the momentum. The optimal momentum formulation validated in the source code is employed in this paper, where it is $m = 1 - 0.5^{1/7000}$. This setting maintains a relatively high running mean in the early training stage, enables low-quality samples to participate in initial learning, and prevents excessively slow model convergence. During initialization $\bar{L}_{IoU}^{(0)} = 1$, and the running mean is updated only during the training phase while remaining fixed in the inference phase.

The anchor box outlier degree β is further defined as the ratio of the current IoU loss to the running mean:

$$\beta = \frac{L_{IoU}}{\bar{L}_{IoU}} \tag{15}$$

A smaller β value indicates a higher degree of overlap between the anchor frame and the ground truth frame and better quality; a larger β value indicates lower-quality anchor frames or outliers with incorrect annotations.

A non-monotonic focus coefficient r is constructed based on outlierness to achieve dynamic allocation of gradient gain:

$$r = \frac{\beta}{\delta \gamma^{\beta - \delta}} \tag{16}$$

where γ and δ are hyperparameters; this paper adopts the optimal values of $\gamma = 1.9$ and $\delta = 3$, which were jointly validated by the original paper and its code. This function reaches its maximum at $\beta = \delta$, meaning that samples of average quality with an outlier score of δ achieve the highest gradient gain. The gradient gain for both high-quality samples ($\beta < \delta$) and low-quality samples ($\beta > \delta$) decreases as the outlier degree deviates from δ, thereby achieving dual suppression of these two sample types. The final WIoUv3 loss is the product of the base IoU loss and a nonmonotonic focus factor:

$$L_{WIoUv3} = rL_{IoU} \tag{17}$$

WIoUv3 is integrated into the YOLOv8s_BSE model to replace the original CIoU bounding box regression loss. For infrared image detection of substation equipment, this loss function can effectively mitigate the interference of labeling noise and improve the localization accuracy of tiny targets with only 3 to 10 pixels, such as miniature connecting components. The improved loss function achieves more precise localization and recognition for power equipment and alleviates the sample imbalance problem to a certain extent.

3. Experiments and Results Analysis

This chapter details the experimental dataset construction, training parameter configuration, and evaluation metric definition. Through comparative experiments with mainstream object detection models, ablation studies on each improved module, and detection result visualization, we systematically validate the superior performance of the YOLOv8s_BSE model for substation equipment infrared detection tasks.

3.1 Experimental Description

1) Dataset: The infrared images of substation equipment used in this study were collected from real-world environments by inspection robots and enhanced using techniques such as image flipping and cropping. The dataset comprises a total of 5,527 images and includes 14 categories of key substation equipment: circuit breakers, current transformers, connectors, connection points, load switches, low-voltage circuit breakers, molded-case circuit breakers, busbars, power capacitors, voltage transformers, vacuum circuit breakers, equipment bodies, connection structures, and core components. See Table 1 for specific category details. An annotation tool was used to annotate the location and recognition information for all images in the dataset. The dataset is divided into three subsets: 3,882 training images, 1,102 validation images, and 543 test images. ocation and recognition information for all images in the dataset. The dataset is divided into three subsets: 3,882 training images, 1,102 validation images, and 543 test images.

Table 1. Dataset Details

	ACB	CT	Conn	CP	LA	LBS	MCCB	MOF	PF	PT	VCB	Body	CS	Core	BG
Total	89	408	203	1314	138	277	618	251	33	284	99	380	690	577	166
Train	62	286	142	920	97	194	433	176	23	199	69	266	483	404	116
Val	18	81	41	263	27	55	123	50	7	57	20	76	138	115	33
Test	9	41	20	131	14	28	62	25	3	28	10	38	69	58	17

2) Experimental Setup and Details: To ensure fair and rigorous comparison, all models were trained and tested under exactly the same experimental conditions.The model used in this paper is built on the PyTorch framework. Experiments were conducted on a Windows 10 system equipped with an NVIDIA GeForce RTX 4090, using CUDA version 11.8 to accelerate training. The input image size was set to 640×640, the training batch size to 32, and the total number of training epochs to 500. The model was optimized using the stochastic gradient descent (AdamW) optimizer, with an initial learning rate of 0.0001, momentum of 0.937, and weight decay of 0.0005. During training, a cosine-smoothed learning rate scheduling strategy was employed, and data augmentation methods such as Mosaic and MixUp were introduced to enhance the model's generalization ability.

3) Evaluation Metrics: This study uses precision (P), recall (R), mean average precision (mAP), floating-point operations (GFLOPs), number of parameters, and inference speed (FPS) as metrics for evaluating model performance. The formulas for precision and recall are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

In the formula, TP (True Positive) represents the number of true positive samples, FP (False Positive) represents the number of false positive samples, and FN (False Negative) represents the number of false negative samples. Negative) represents the number of false negative samples. Average Precision (AP) is calculated as the area under the precision-recall curve, while mAP is the average of the AP values across all categories. This study primarily uses the mAP@0.5 (AP at an F1 score threshold of 0.5) and the mAP@0.5:0.95 (AP across F1 score thresholds ranging from 0.5 to 0.95) as core accuracy metrics. The number of floating-point operations and the number of parameters are used to evaluate the computational complexity and lightweight nature of the model, while inference speed (FPS) is used to assess the model's real-time performance.

3.2 Model Performance and Comparative Analysis

3.2.1 Classification Performance and Confusion Matrix Analysis

To evaluate the classification performance of the YOLOv8s_BSE model, a confusion matrix was generated based on the test set, with the results shown in Figure 5.

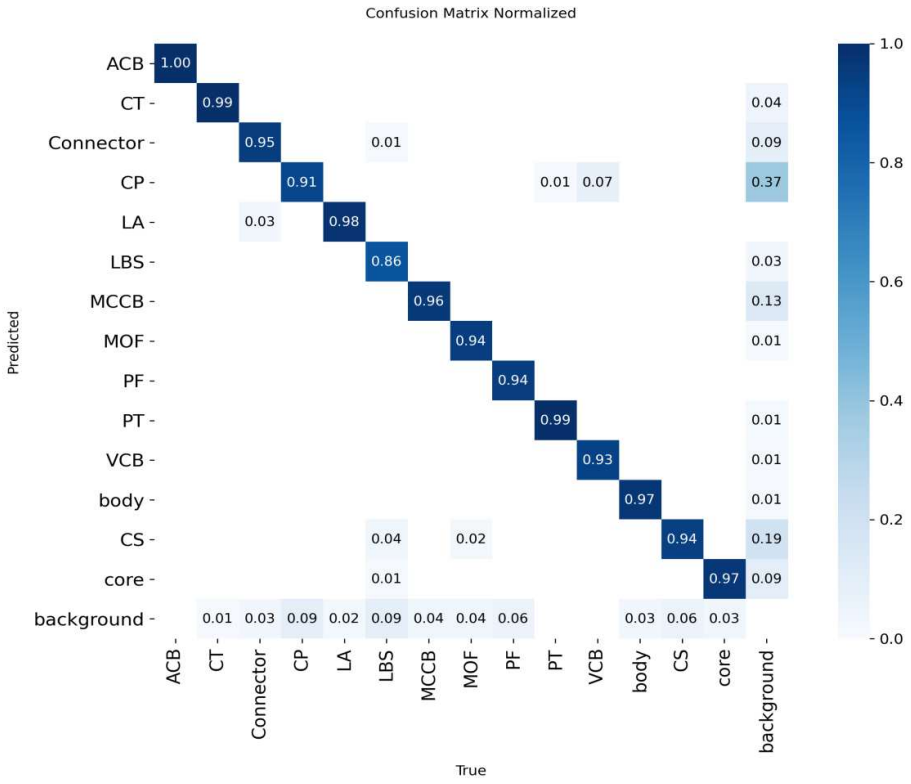


Figure 5. Confusion Matrix

The experimental results indicate that the model achieved an overall classification accuracy of 95.8%, with high classification performance across all equipment categories. In terms of classification performance across categories, the classification accuracy rates for the six sample categories—circuit breakers, voltage transformers, vacuum circuit breakers, equipment bodies, connection structures, and core components—reached 100%, 99%, 93%, 97%, 94%, and 97%, respectively, indicating that the model can accurately capture the unique distinguishing features of these categories. The classification accuracy for the five sample categories—current transformers, connecting components, load switches, busbars, and molded-case circuit breakers—ranged between 94% and 99%, demonstrating the model's generalization robustness in multi-class scenarios. The classification accuracy for the two sample categories—power capacitors and low-voltage circuit breakers—was relatively lower at 91% and 86%, respectively, yet still fell within an acceptable high-performance range.

Error analysis shows that the model's primary error types are false positives and false negatives. False positives mainly stem from similar thermal distribution patterns among functionally similar devices. For example, device body samples have a 4% misclassification rate as circuit breakers and a 2% rate as current transformers. These devices exhibit similar temperature gradients during normal operation, making them difficult to distinguish based on single features in infrared images. False

negatives mainly occur with small targets and devices with homogeneous thermal signatures. For example, connector samples have a 5% miss rate or misclassification rate as circuit breakers or current transformers. Connectors are only 3–10 pixels in size with weak thermal signatures, making them easily obscured by complex background noise. Additionally, current and voltage transformers share similar structural configurations, causing confusion under low-contrast conditions.

In summary, the YOLOv8s_BSE model demonstrates excellent robustness in the multi-class substation equipment infrared classification task. It achieves near-perfect accuracy for medium-to-large targets and significantly improved performance for small targets, with only a small number of errors occurring in a few categories with indistinct thermal features. This classification performance provides a solid foundation for subsequent object detection and precise localization.

3.2.2 Performance Comparison of Mainstream Detection Models

To comprehensively validate the overall performance of the YOLOv8s_BSE model in substation equipment infrared detection tasks, we selected three representative lightweight detection models as benchmarks for comparative analysis. The first category includes the latest lightweight models from the YOLO series: YOLOv6s, YOLOv8s, YOLOv9s, YOLOv10s, YOLOv11s, and YOLOv12s. The second category includes industrial-grade and state-of-the-art baseline models with non-YOLO architectures: the traditional single-stage lightweight model SSD-MobileNetV2 and the Transformer-based real-time detection model RT-DETR-L. The third category includes YOLOv8s variant models with single improved modules, designed to quantify the independent contribution of each innovation. We trained and evaluated all models using the same dataset, training strategy, and testing environment. The comparison results are presented in Table 2.

Table 2. Performance Comparison of Different Models

Models	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95	GFLOPs	Params/M	FPS
SSD-MobileNetV2	91.1	72.1	91.1	0.641	2.806	3.225	31
RT-DETR-L	90.5	91.7	94.2	0.715	24.3	32.8	252.96
YOLOv6s	93.5	92.0	94.9	0.748	13.1	4.50	200.11
YOLOv8s	93.9	93.9	95.9	0.764	28.7	11.14	206.35
YOLOv9s	91.9	93.3	95.4	0.759	27.4	7.29	145.18
YOLOv10s	93.5	93.5	95.9	0.767	25.1	8.13	295.14
YOLO11s	94.1	93.6	96.2	0.765	21.7	9.46	197.45
YOLO12s	94.6	90.8	94.5	0.752	21.7	9.28	172.67
YOLOv8s_GSConv	93.8	93.3	95.6	0.761	27.9	11.39	199.04
YOLOv8s_Concat_BiFPN	95.7	92.4	95.7	0.759	28.7	11.14	213.88
YOLOv8s_EMA	95.4	94.1	96.4	0.761	29.8	11.21	202.81
YOLOv8s_BSE	95.8	94.7	96.8	0.761	26.5	9.61	269.27

Note: The high FPS of RT-DETR-L is due to its efficient end-to-end architecture that eliminates the need for non-maximum suppression (NMS), which gives it a speed advantage in pure forward pass testing.

It can be observed from the results in Table 2 that YOLOv8s_BSE achieves the optimal trade-off between accuracy and efficiency among all comparative models.

1) Compared with baseline models based on non-YOLO architectures, the proposed model achieves significant comprehensive performance advantages over non-YOLO architecture baselines. Compared with the traditional industrial lightweight model SSD-MobileNetV2, YOLOv8s_BSE improves mAP@0.5 by 5.7 percentage points and increases inference speed by 7.7 times, effectively solving the accuracy deficiency problem of traditional lightweight models. Compared with the state-of-the-art Transformer-based RT-DETR-L, the proposed model has only 29.3% of the parameter count of RT-DETR-L and reduces GFLOPs by 9.1%, while improving mAP@0.5 by 2.6 percentage points. This makes it more suitable for embedded deployment in resource-constrained substation environments.

2) Compared with the latest lightweight models in the YOLO series, YOLOv8s_BSE also achieves significant performance advantages over the latest YOLO series lightweight models. Compared with the baseline YOLOv8s, the proposed model improves precision, recall, and mAP@0.5 by 1.9, 0.8, and 0.9 percentage points, respectively. Meanwhile, it reduces GFLOPs by 7.7%, decreases parameter count by 13.7%, and increases inference speed by 30.5%. Even compared with YOLOv11s, which achieves the best overall performance in the current YOLO series, the proposed model still outperforms it by 0.6 percentage points in mAP@0.5 and achieves 36.4% higher inference speed.

3) In terms of the comparative results of individual improved modules, all innovative components contribute positively to the overall model performance. Among them, the EMA attention module contributes the most to accuracy improvement, yielding a 0.5 percentage point gain in mAP@0.5 when applied alone. The Concat-BiFPN module further raises the model recall by enhancing multi-scale feature fusion. The C2f-Star module realizes model lightweight compression while maintaining

almost no accuracy loss. The synergistic effect of the four modules enables YOLOv8s_BSE to maintain high detection accuracy and simultaneously satisfy the strict requirements on inference speed and deployment cost for real-time substation inspection.

The precision–recall curve, together with the mAP@0.5 and mAP@0.5:0.95 curves during the training process, is presented in Figure 6.

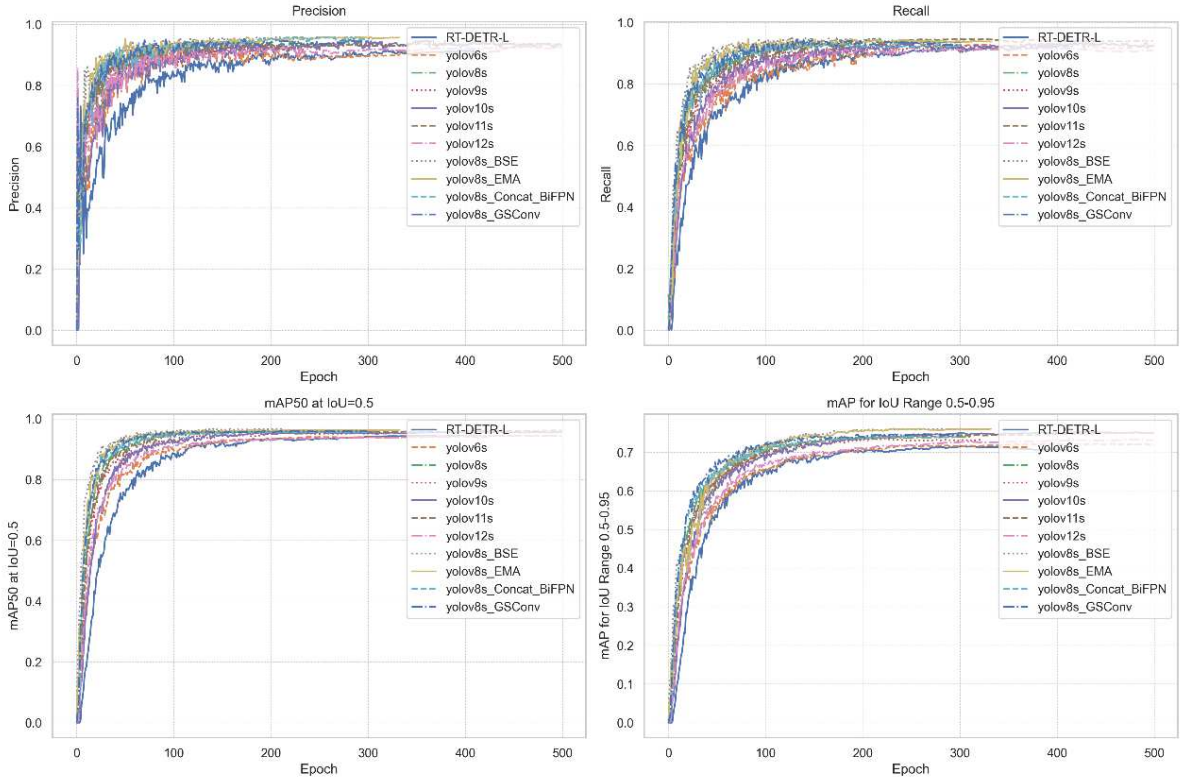


Figure 6. Precision–recall curves, mAP@0.5, and mAP@0.5:0.95 comparison curves of different models

As can be seen from the trend curves in Figure 6, all models exhibit rapid performance improvement during the initial 0–100 training epochs, after which the convergence rate slows down and the performance gradually stabilizes. YOLOv8s_BSE begins to significantly outperform the baseline YOLOv8s in all evaluation metrics after approximately 60 training epochs. When trained for 500 epochs, the model reaches its peak performance with no obvious fluctuations in the curves, indicating that it possesses superior feature learning capability and training stability. In contrast, RT-DETR-L exhibits a slower convergence rate in the early training phase, and its performance only gradually catches up with some YOLO-series models after 200 epochs; however, its final mAP@0.5 remains lower than that of the proposed model. All variants with individual improved modules exhibit faster convergence rates than the baseline YOLOv8s, further verifying the optimization effect of each innovation on the model training process.

3.3 Ablation Experiments on Module Effectiveness

To systematically verify the independent contribution of each improved module and the actual detection performance of the model, ablation experiments are conducted in this section to quantify the performance impact of each module. Meanwhile, visualization comparisons under typical working conditions are performed to intuitively demonstrate the detection superiority of YOLOv8s_BSE in complex infrared scenarios of substation equipment.

3.3.1 Ablation Analysis of Module Effectiveness

To quantitatively evaluate the contribution of each improved module to model performance, ablation experiments are carried out under identical experimental conditions, and the results are listed in Table 3.

Experimental results demonstrate that all four improved modules yield positive contributions to model performance with remarkable synergistic effects. The ablation experiments are designed in a progressive stacking manner, which can clearly quantify the independent contribution of each module and verify the rationality of the overall architecture design.

Table 3. Ablation Experiments on Module Effectiveness

Models	WIoU	EMA	C2f_Star	Concat_BiFPN	P/%	R/%	mAP@50/%	GFLOPs/G
YOLOv8s	×	×	×	×	93.9	93.9	95.9	28.7
YOLOv8s-W	✓	×	×	×	95.1	94.1	96.4	28.7
YOLOv8s-WE	✓	✓	×	×	95.4	94.1	96.4	29.8
YOLOv8s-WES	✓	✓	✓	×	95.5	93.8	96.1	26.5
YOLOv8s_BSE	✓	✓	✓	✓	95.8	94.7	96.8	26.5

First, introducing only the WIoU loss function (YOLOv8s-W) increases the model's mAP@0.5 by 0.5 percentage points without incurring any additional computational overhead. This verifies that the dynamic non-monotonic focusing mechanism can effectively improve bounding box regression accuracy by adaptively adjusting gradient weights for samples of different qualities. Subsequent integration of the EMA attention module (YOLOv8s-WE) further elevates the detection accuracy to 95.4% at the cost of a marginal increase in computational complexity to 29.8 GFLOPs, indicating that feature enhancement is achieved with minimal computational consumption.

Subsequently, the introduction of the C2f-Star module (YOLOv8s-WES) reduces the GFLOPs to 26.5, achieving a prominent lightweight effect. Although the recall decreases slightly due to the reduction in model parameters, the overall performance still remains superior to the baseline YOLOv8s model. This trade-off between accuracy and computational complexity is highly valuable for embedded deployment scenarios in substations with limited computing resources.

Finally, incorporating the Concat-BiFPN module (YOLOv8s_BSE) boosts recall and mAP@0.5 by 0.9 and 0.7 percentage points respectively while preserving the lightweight characteristics. This effectively compensates for the minor accuracy loss introduced by the C2f-Star module and realizes the full synergistic potential of all proposed improvements, achieving the optimal balance between detection accuracy and inference efficiency.

In summary, the WIoU loss function delivers the most prominent performance gain as a single module, the C2f_Star module achieves the core lightweight design goal, and the Concat-BiFPN module acts as a key component for performance improvement by enhancing multi-scale feature fusion. The synergistic combination of the four modules enables the YOLOv8s_BSE model to achieve the highest detection accuracy while maintaining the lowest computational complexity.

The WIoUv3 loss function is adopted in this paper because it has been proven to outperform CIoU, SIoU and EIoU in most recent target detection tasks [27], especially for low-quality samples. This paper focuses on verifying the effectiveness of combining WIoUv3 with the proposed feature extraction and fusion modules. Similarly, this paper compares the proposed Concat-BiFPN with the original PANet structure used in YOLOv8s. A more comprehensive comparison of different loss functions and a direct comparison between Concat-BiFPN and the original BiFPN will be conducted in future work.

3.3.2 Detection Result Visualization and Comparison

To intuitively verify the detection robustness of YOLOv8s_BSE in complex infrared scenarios of substation equipment, representative images of seven typical working conditions from the test set are selected for horizontal comparison. These working conditions cover the core challenges in equipment infrared detection, including low contrast, dense small targets, occlusion, mixed multi-scale targets, local overheating, overexposure under large temperature differences, and inclined targets, which can fully reflect the model's performance under diverse, complex conditions. The comparison models include YOLOv8s, YOLOv10s, YOLOv11s, and SSD-MobileNetV2, and the detection results are illustrated in Fig. 7-11.



a b c d e

Figure 7. Detection results comparison of different models on substation infrared images. (a) YOLOv8s; (b) YOLOv10s; (c) YOLOv8s_BSE; (d) YOLOv11s; (e) SSD-MobileNetV2

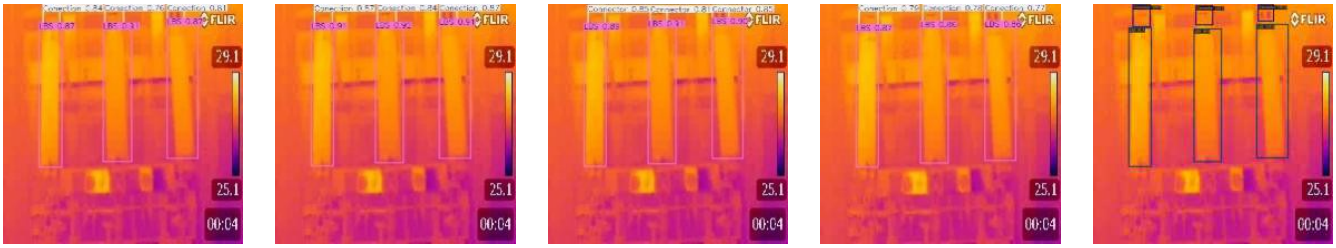
As shown in Figure 7, in the insulator occlusion scenario, the target is obscured by multiple layers of insulators and power lines, with less than 40% of the target visible. SSD-MobileNetV2 misclassifies background wires as equipment; the detection confidence of YOLOv8s and YOLOv11s drops below 0.6, while YOLOv10s fails to detect the target. YOLOv8s_BSE, through the complementary fusion of multi-scale features, infers the complete target using locally visible information, maintaining a confidence level above 0.85 with no false positives or missed detections.



a b c d e

Figure 8. Detection results comparison of different models on substation infrared images. (a) YOLOv8s; (b) YOLOv10s; (c) YOLOv8s_BSE; (d) YOLOv11s; (e) SSD-MobileNetV2

As shown in Figure 8, in a multi-scale mixed-object scene, the image simultaneously contains a large cabinet, a medium-sized current transformer, and a tiny connector, with a target scale range of 1:20. SSD-MobileNetV2 fails to recognize small targets; YOLOv8s and YOLOv11s exhibit detection confidence below 0.7 for connectors, while YOLOv10s shows slight bounding box misalignment. YOLOv8s_BSE achieves precise localization for targets of all scales, with confidence consistently above 0.9, validating the adaptability of the Concat_BiFPN bidirectional feature propagation mechanism.



a b c d e

Figure 9. Detection results comparison of different models on substation infrared images. (a) YOLOv8s; (b) YOLOv10s; (c) YOLOv8s_BSE; (d) YOLOv11s; (e) SSD-MobileNetV2

As shown in Figure 9, in a dense scene with multiple connectors, the image contains numerous tiny targets ranging from 3 to 10 pixels in size, with severe overlap of thermal features between devices. SSD-MobileNetV2 exhibited extensive missed detections, while YOLOv8s and YOLOv11s each missed two targets, and YOLOv10s missed one. YOLOv8s_BSE successfully identified all targets with precise bounding box localization and detection confidence scores all above 0.8. These results demonstrate the ability of the C2f_Star module and Concat_BiFPN to enhance weak thermal features, effectively resolving the issue of missed small targets.



a b c d e

Figure 10. Detection results comparison of different models on substation infrared images. (a) YOLOv8s; (b) YOLOv10s; (c) YOLOv8s_BSE; (d) YOLOv11s; (e) SSD-MobileNetV2

As shown in Figure 10, in the tilted target scenario, the device is tilted at a 30° angle due to the shooting perspective, causing traditional models to exhibit positioning errors. The angle between the bounding boxes of SSD-MobileNetV2 and the target exceeds 20°, while that of YOLOv8s and YOLOv11s exceeds 10°, and that of YOLOv10s is approximately 8°. YOLOv8s_BSE enhances global feature representation through the EMA attention mechanism, enabling it to adapt to changes in target orientation. The angle between its bounding boxes and the target is less than 3°, resulting in significantly higher localization accuracy than the comparison models.



a b c d e

Figure 11. Detection results comparison of different models on substation infrared images. (a) YOLOv8s; (b) YOLOv10s; (c) YOLOv8s_BSE; (d) YOLOv11s; (e) SSD-MobileNetV2

As shown in Figure 11, in high-temperature-difference overexposure scenarios, where the temperature difference between the device and the background exceeds 40°C, local areas become overexposed, and heat distribution is extremely uneven. SSD-MobileNetV2 exhibits severe scaling errors in its bounding boxes, while YOLOv8s and YOLOv11s show positioning shifts exceeding 15°, and YOLOv10s has a confidence score below 0.7. YOLOv8s_BSE's bounding boxes closely align with the object contours, and its confidence scores remain stable, unaffected by overexposed areas, demonstrating the model's robustness under extreme temperature conditions.

Overall, YOLOv8s_BSE exhibits the best detection performance across all seven typical scenarios. It delivers more precise bounding box localization, higher detection confidence, and stronger robustness against small targets, occlusions, low contrast interference, and tilted targets. This is consistent with the quantitative results presented in the previous section, where the mAP@0.5 is improved by 0.9 percentage points. Combined with the 4.2 percentage point gain in small-target detection accuracy, this further validates the effectiveness and engineering practicability of the proposed method.

4. Conclusion

To address the persistent challenges of low contrast, strong noise interference, and blurred small-target features in infrared detection of substation equipment, we propose a lightweight multi-scale fusion detection model, YOLOv8s-BSE. Through collaborative optimization of backbone feature enhancement, neck multi-scale fusion, and detection head loss function, we construct a progressive "feature enhancement-multi-scale fusion-precise localization" architecture, achieving a favorable balance between detection accuracy and inference efficiency.

Extensive experiments confirm that the proposed model achieves 96.8% mAP@0.5 on our self-built 14-category substation infrared dataset, a 0.9 percentage point gain over the baseline YOLOv8s, and concurrently reduces parameters to 9.61 M, cuts computational complexity by 7.7%, and reaches an inference speed of 269.27 FPS. This lightweight design exhibits promising deployment potential: its parameter scale and computational complexity are well within the operational limits of mainstream edge computing chips, establishing a solid technical foundation for deployment on inspection robots, unmanned aerial vehicles (UAVs), and edge nodes, and is expected to facilitate the intelligent transformation of substation operation and maintenance.

Despite these promising results, several limitations of this work should be acknowledged. First, the dataset lacks samples under extreme weather conditions, which may constrain the model's generalization capability in complex field environments. Second, detection performance for ultra-small targets (<3 pixels) and samples with temperature differences below 2°C still requires further improvement. Third, inference speed is only tested on a high-performance GPU platform, and end-to-end deployment performance on edge devices such as Jetson Xavier NX has not yet been fully validated. Finally, the model only detects thermal anomalies without classifying fault severity.

Future research directions will focus on expanding the dataset to cover diverse environmental conditions, designing dedicated ultra-small target detection heads, conducting model quantization and edge deployment testing, and integrating visible-infrared multi-modal fusion and fault severity classification functions to provide more comprehensive technical support for intelligent power grid operation and maintenance.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This work was supported by Jilin Province Science and Technology Development Plan Project of Grant No. YDZJ202401367ZYTS.

Data Availability

The datasets generated and analysed during the current study are not publicly available due to the confidentiality requirements of the power grid company and national security regulations for power grid infrastructure but are available from the corresponding author on reasonable request.

References

1. Han X, Guo J, Pu T, et al. Theoretical foundation and directions of electric power artificial intelligence (I): hypothesis analysis and application paradigm[J]. Pro. CSEE, 2023, 8: 2877-2891.
2. International Energy Agency. International Energy Agency (IEA) World Energy Outlook 2022[J]. International Information Administration, 2022, 524.
3. Zeng Z P. On the maintenance and common fault handling methods of substation operating equipment[J]. China Plant Engineering, 2024, (5): 53-55.
4. Liu Y, Pei S, Wu J, et al. Deep learning based target detection method for abnormal hot spots infrared images of transmission and transformation equipment[J]. Southern Power System Technology, 2019, 13(2): 27-33.
5. Junhuang Z, Tingcheng H, Xiaoyu X I E, et al. Review of application research of video image intelligent recognition technology in power transmission and distribution systems[J]. Electric Power, 2021, 54(1): 124.
6. Jin L, Kim D, Abu-Siada A. State-of-the-art review on asset management methodologies for oil-immersed power transformers[J]. Electric Power Systems Research, 2023, 218: 109194.
7. Wu C, Wu Y, He X. Infrared image target detection for substation electrical equipment based on improved faster region-based convolutional neural network algorithm[J]. Review of Scientific Instruments, 2024, 95(4).
8. Ou J, Wang J, Xue J, et al. Infrared image target detection of substation electrical equipment using an improved faster R-CNN[J]. IEEE Transactions on Power Delivery, 2022, 38(1): 387-396.
9. Liu T, Li G, Gao Y. Fault diagnosis method of substation equipment based on You Only Look Once algorithm and infrared imaging[J]. Energy Reports, 2022, 8: 171-180.
10. Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11534-11542.
11. Zhao Z B, Feng S, Zhao W Q, et al. A thermal image detection method for substation equipment incorporating knowledge migration and improved YOLOv6[J]. CAAI Trans. Intell. Syst, 2023, 6: 1213-1222.
12. Changzheng D, Mingze L I U, Tian F U, et al. Infrared Image Recognition of Substation Equipment Based on Improved YOLOv7-Tiny Algorithm[J]. Infrared technology, 2025, 47(1): 44-51.
13. Lu L, Li M L, Xiong W, et al. Infrared image detection of substation equipment based on improved YOLOv8[J]. Infrared Technol, 2024, 46: 1-7.
14. Yang T, Ling D, Shi Q, et al. YOLOv8-MCDE for lightweight detection of small instruments in complex backgrounds from inspection robots' perspective[J]. Scientific Reports, 2025, 15(1): 32060.
15. Sun Y, Sun X, Lin Y, et al. Substation equipment defect detection based on improved yolov8[J]. Sensors, 2025, 25(11): 3410.
16. Shan D, Yang S, Zhao Z, et al. Lightweight infrared object detection network based on improved ssd[C]//Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence. 2023: 347-352.
17. Zhai X, Huang Z, Li T, et al. YOLO-Drone: an optimized YOLOv8 network for tiny UAV object detection[J]. Electronics, 2023, 12(17): 3664.
18. Huangfu Z, Li S. Lightweight you only look once v8: An upgraded you only look once v8 algorithm for small object identification in unmanned aerial vehicle images[J]. Applied Sciences, 2023, 13(22): 12369.
19. Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

20. Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

21. Babiloni F, Marras I, Kokkinos F, et al. Poly-nl: Linear complexity non-local layers with 3rd order polynomials[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10518-10528.

22. Muslikhin M, Nasuha A, Arifin F, et al. Troop camouflage detection based on deep action learning[J]. IAES International Journal of Artificial Intelligence, 2022, 11(3): 859.

23. Wu X, Qian S, Yang M. Detection of safety helmet-wearing based on the YOLO_CA model[J]. Computers, Materials, & Continua, 2023, 77(3): 3349.

24. Li Y, Leng Q, Lu J. YOLOv8n-EMA-ODC wafer defect classification model[C]//Seventh International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2024). SPIE, 2024, 13229: 1322902.

25. Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2023: 1-5.

26. Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

27. Tong Z, Chen Y, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint arXiv:2301.10051, 2023.

28. Jocher G, Chaurasia A, Qiu J, et al. Ultralytics YOLO: A unified framework for real-time object detection, segmentation, and classification. Ultralytics, 2026. https://github.com/ultralytics/ultralytics (Version 8.4.25, accessed 23 March 2026)